Comparison of Conceptual Graphs

Manuel Montes-y-Gómez,
Alexander Gelbukh,
Aurelio López-López

In intelligent knowledge-based systems, the task of approximate matching of knowledge elements has crucial importance. We present the algorithm of comparison of knowledge elements represented with conceptual graphs. The method is based on well-known strategies of text comparison, such as Dice coefficient, with new elements introduced due to the bipartite nature of the conceptual graphs. Examples of comparison of two pieces of knowledge are presented. The method can be used in both semantic processing in natural language interfaces and for reasoning with approximate associations.

1 INTRODUCTION

For an intelligent knowledge-based system, it is important to be able to approximately compare two pieces of knowledge, answering the questions: How similar are the two situations? What situations in the knowledge base are similar to the given one? What pieces of knowledge could be useful for reasoning with the given one? This is similar to the behavior of a person who has just learned the piece of news that John came late to the party. The person recalls the similar pieces of knowledge (s)he already knows: Last week John came late to the class, or Jack came to the party too. Also, the person can generalize the available knowledge: Boys like to attend parties. An intelligent system should be able to model this behavior.

For this, the system should be able to compare pieces of knowledge in a quantitative manner rather than on the equal-or-not basis. The task of recalling "similar" knowledge and generalizing the available knowledge in an intelligent agent are similar to the tasks of natural language processing involving approximate matching, such as information retrieval, text mining, and abstracting. These tasks were our main motivation in this research.

For plain keyword set representation of text, like {*algorithm*, *binary*, *search*}, many different similarity measures are proposed, for instance, the Dice coefficient, the Jaccard coefficient, the Cosine coefficient [Rasmussen 1992], etc. For the representation with binary term weights, the Dice coefficient is calculated as follows:

$$S_{D_1,D_2} = \frac{2n(D_1 \cap D_2)}{n(D_1) + n(D_2)} \quad , \tag{1}$$

where $n(D_i)$ is the number of terms in D_i, and $n(D_1 \cap D_2)$ is the number of terms that the two documents D_i and D_j have in common. Because of its simplicity and normalization, we take it as the basis for the similarity measure we propose.

Conceptual graphs have been used as a representation of text contents [Myaeng 1990; Genest & Chein 1997]. In the field of information retrieval, different similarity measures have been described for comparing the query graph with the graphs from the knowledge base. The matching criterion most widely used for conceptual graphs is that if the query graph is completely contained in the given graph, then the given graph is relevant for (i.e., matches with) the given query graph. This criterion means that the contents of the found piece of information have to be more particular than the query piece.

We propose a more flexible quantitative approximate matching criterion. First, we introduce the notion of the conceptual graph and describe the process of transformation of a text to a set of conceptual graphs. Then, we explain the main idea of the comparison of two conceptual graphs, and give the corresponding formulae. Finally, we give some examples of comparison of conceptual graphs.

2 SIMILARITY MEASURE

Given two conceptual graphs G_1 and G_2 respectively and the graph $G_1 \cap G_2 = G_c$, we define the similarity s between them as a combination of two values: their conceptual similarity s_c and their relational similarity s_r.

The conceptual similarity s_c expresses how many concepts the two graphs G_1 and G_2 have in common. We calculate it using an expression analogous to the well-known Dice coefficient used in information retrieval (Rasmussen 1992):

$$s_c = \frac{2n(G_c)}{n(G_1) + n(G_2)} \tag{2}$$

where $n(G)$ is the number of concept nodes of a graph G. This expression varies from 0 (when the two graphs have no concepts in common) to 1 (when the two graphs consist of the same set of concepts).

The relational similarity s_r indicates how similar the relations between the same concepts in both graphs are, that is, how similar the information about these concepts contained in the two pieces of knowledge is. In a way, it shows how similar the contexts of the common topics in both graphs are.

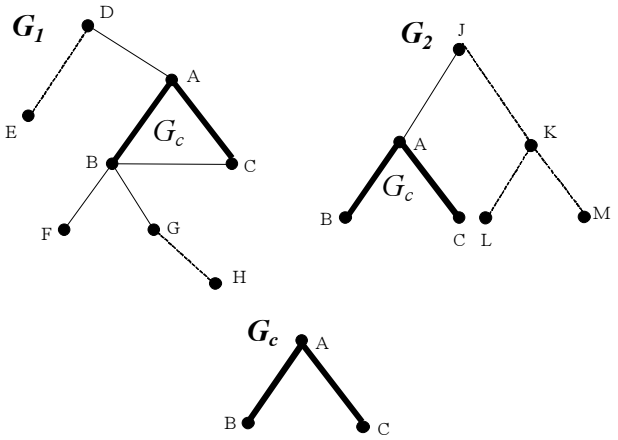

Figure 1. Intersection of the two graphs.

2.1 Main equation

We define the relational similarity s_r to measure the proportion between the degree of connection of the concept nodes in G_c, on the one hand, and the degree of connection of the same concept nodes in the original graphs G_1 and G_2, on the other hand. With this idea, a relation between two concept nodes conveys less information about the context of these concepts if they are highly connected in the original graphs, and conveys more information when they are weakly connected in the original graphs. We formalize this using a modified formula for the Dice coefficient:

$$s_r = \frac{2m(G_c)}{m_{G_c}(G_1) + m_{G_c}(G_2)} \tag{3}$$

where $m(G_c)$ is the number of the arcs (the relation nodes in the case of conceptual graphs) in the graph G_c, and $m_{G_c}(G)$ is the number of the arcs in the immediate neighborhood of the graph G_c in the graph G. The immediate neighborhood of $G_c \subseteq G$ in G consists of the arcs of G with at least one end belonging to G_c.

2.2 Comparison of the graphs

Figure 2 illustrates these measures. In this figure, the nodes *A*, *B*. and *C* are the conceptual nodes common for G_1 and G_2 and thus belonging to G_c. Bold lines represent the arcs (relation nodes) common to the two graphs. The arcs marked with the symbol ✓ constitute the immediate neighborhood of the graph G_c (high-

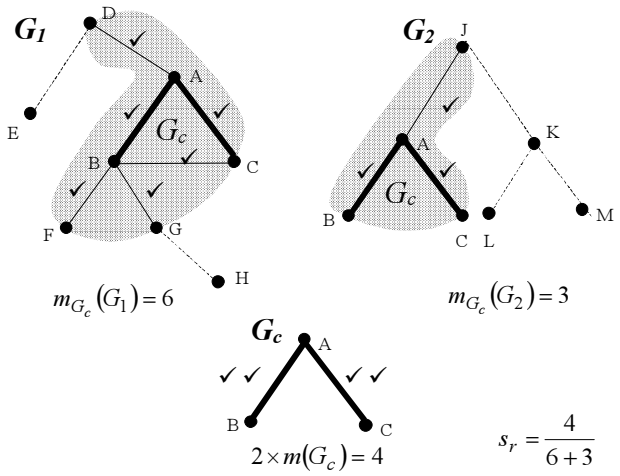

Figure 2. Calculation of relational similarity.

lighted areas), their number is expressed by the term $m_{G_c}(G_i)$ in the formula above.

2.2.1 The formula of comparison

The value of $m_H(G)$ for a subgraph $H \subseteq G$ in practice can be calculated as follows:

$$m_H(G) = \sum_{c \in H} deg_G c - m(H) \quad , \tag{4}$$

where $deg_G c$ is the degree of concept node c in the graph G, i.e., the number of the relation nodes connected to the concept node c in the graph G, and $m(H)$ is the number of relation nodes in the graph H.

2.2.1.1 A 4-level title

Using too deep levels of titles is not recommended (though not prohibited).

3 CONCLUSIONS

We have described a method for measuring the similarity between two conceptual graphs representing two pieces of knowledge in an intelligent system. The method is based on the idea of the Dice coefficient, a widely used measure of

similarity for the keyword representations of texts. It also incorporates some new characteristics derived from the conceptual graph structure, for instance, the combination of two complementary sources of similarity: the conceptual similarity and the relational similarity.

The method of comparison of conceptual graphs has potential uses not only in intelligent agents and knowledge bases, but also in other tasks of knowledge management, such as information retrieval systems, text mining, and document classification (actually, this was our main motivation and the area of experiments).

REFERENCES

Genest D., and M. Chein, 1997. *An Experiment in Document Retrieval Using Conceptual Graphs*. Conceptual structures: Fulfilling Peirce´s Dream. Lecture Notes in artificial Intelligence 1257. August 1997.

Myaeng, Sung H., 1990. *Conceptual Graph Matching as a Plausible Inference Technique for Text Retrieval*. Proc. of the 5th Conceptual Structures Workshop, held in conjunction with AAAI-90, Boston, Ma, 1990.

Rasmussen, Edie, 1992. *Clustering Algorithms*. Information Retrieval: Data Structures & Algorithms. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall, 1992.

Sowa, John F., 1983. *Conceptual Structures: Information Processing in Mind and Machine*. Ed. Addison-Wesley, 1983.

Manuel Montes-y-Gómez is a doctoral student at the Center for Computing Research (CIC), National Polytechnic Institute (IPN), Av. Juan de Dios Bátiz, esq. Mendizabal, Zacatenco, 07738, Mexico D.F., Mexico. He can be reached at mmontesg@susu.inaoep.mx.

Alexander Gelbukh is the head of the Laboratory of Natural Language Processing of the same Institute, author of about 80 publications in the field of computational linguistics, member of the National System of Researchers (SNI) of Mexico. He can be reached at gelbukh@cic.ipn.mx, see http://www.pollux.cic.ipn/~gelbukh.

Aurelio López-López is a researcher and professor at INAOE, Electrónica. Luis Enrique Erro No. 1 Tonantzintla, Puebla, 72840 México. He can be reached at allopez@gisc1.inaoep.mx, tel. (52 22) 472011, fax (52 22) 470517.

The work was done under partial support of CONACyT, COFAA-IPN, and SNI, Mexico.